Essay on Information extraction based on three articles

I have chosen the following three articles to have talk on:

1. A Biomedical Information Extraction Primer for NLP Researchers
2. Lithium NLP: A System for Rich Information Extraction from Noisy User Generated Text on Social Media
3. Large Scale Relation Extraction with Reinforcement Learning

Nowadays, neural networks are entering our daily lives closer and closer. We don’t even pay attention to that, but it is in our daily routines like social media apps, market places, search engines, video apps and so on. All the commercial companies and academia people trying to find the best benefit-cost effective solutions. Nevertheless, there are still a lot of obstacles that we are going to cover by examining three articles above and describe possible solutions that the authors might have found.

The main three categories that we might consider when talking about Information Extraction in general. Those are name entity recognition, relation extraction and event extraction. As we go deeper there are a lot of obstacles and challenges to identify the best possible outcomes without human interaction, since human interaction is only possible on limited data and also cost effective, thus giving better results versus current models, which are evolving from day to day. Let’s deep dive into the articles and see what solutions do they propose to any of the categories on Information Extraction mentioned above.

So, the first article that is on the list is about biomedical information extraction. This paper explains all the challenges in all three categories on Information Extraction. Name entity recognition (NER) in the biomedical domain is a complex problem. First of all, there are many synonymy words that are used in biomedicine. For instance, cyclin-dependent kinase inhibitor p27 and p27kip1 are the same names for proteins. The second issue is that there are different semantic meanings for the same abbreviations like RA, which can mean either right atrium or rheumatoid arthritis, which is quite challenging even for human being who is not in the knowledge of the topic to identify the right meaning for any abbreviation in medical context. The third challenge is that entity names are subject to many variants and also tend to change over the time. In order to apply a possible solution to all the challenges above there are several solutions like usage of POS tags for lists with multiple rules depending on simple word features. Also, there is a machine learning approach that uses a semi-conditional random field (semi-CRF) to output labels on all tokens in the sentence and other methods that are still evolving upon our days.

As per Relation Extraction is even harder problem to solve, since it requires to find related entities of many different kinds. Some of these might include protein-protein interactions, disease-gene relations and drug-drug interactions. Due huge amount of biomedical literature it is almost impossible for any human being to find all those correlations. Automatic extraction of relations can assist in forming database to ease the job for medical researcher and find what they need indeed. As an example, let’s take a case for drug-drug interaction database, which can come in handy for suggesting potentially related articles on drug to drug interactions for those doctors and clinicians who are administer multiple drugs simultaneously to their patients. It very important to have an information on how one drug will have an adverse effect on the other. Partial solutions are presented in the article like the dictionary of interacting proteins (DIP) from abstracts in Medline. Another researchers introduced Pointwise Mutual Information (PMI) for corpus level statistics to determine whether a pair of proteins occur together by a chance or because they are actually interact. Other dictionaries were proposed for different relations like Drug-Drug Interaction (DDI), which simply gives the same solution for drug correlations. Using all those dictionaries helps to improve the accuracy on acquired data to use it meaningfully.

Event extraction goes beyond Relation extraction in biomedical context, which refers to change in state of biological molecules such as proteins and DNA. Also, each event type can have a multiple arguments that need to be detected as well. BioNLP’09 helps to find those events and involves prediction of trigger words over 9 event types such as expression, transcription, catabolism, binding and etc. resulting in a graph representation of the possible outcomes. F-score is low and around 50%, but there is a lot of work to do to gain better results.

Summing up the first paper, in general it explained us how complex is biomedical information extraction and how complex it is in order to get the best possible outcomes. Nevertheless, it is quite exciting and challenging task to perform for NLP researchers that demand application of state-of-the-art methods. As deep learning evolves, hand crafted methods will be used less and models such as Convolutional Neural Networks and Long Short-Term Memory methods will take place and help clinicians and researchers in Biomedical domain.

Second article explains Lithium Natural Languages processing system. Lithium NLP extracts a rich set of information such as entities, topics, hashtags and sentiment from the text. Let’s first discuss that complexity that we encounter on social media space. On social media there are noisy un-normalized data, which is less formal and uses a lot of jargons; multi-lingual content that bring even more challenges in acquiring the context out of the data on social media; large scale datasets is another issue, which might consist of millions of documents and state-of-the-art NLP systems are supposed to solve those scales in finite amount of time; rich set of information also a problem that we encounter and cost-efficiency should be in place. In the paper the authors also explain what Knowledge Base they are using for golden benchmark and how they identify the metrics to express the best possible output. They give scores for co-occurrences and based on that correlate the context for the data given. For instance, Michael Jordan is mostly occurred with the words like basketball, NBA and the other Michael Jordan who is professor has entities like ML, AI and so on, so meaning that the same input might have different co-occurrence rate, thus resulting in different outputs in context. They also introduced other metrics like entity importance, topic parents and topic hashtags to better provide the result by semantic.

Now, let’s overview, the Lithium NLP pipeline and split it by stages. Text comes into the system, language is detected using naïve Bayesian filters, which works perfectly on 49 languages. Then text is normalized excluding all unnecessary characters. Then text is broken into sentences, tokenizer is applied. On top of it entity extraction is applied using mention entity co-occurrence dictionary generated offline. After this stage there is an entity disambiguation and linking for abbreviation and other words that might change the meaning of the context, for instance CEO is changed to Chief Executive Office giving a full context to abbreviation. Then topic is projected based on knowledge base. Based on topic projection the hashtags are recommended. The next step is to identify the sentiment of the text, whether it is positive, negative or neutral. And the last stage is to give entity metadata decoration, where the system classifies whether each entity’s type is person, organization, location, film, event, book and so on and provides its location (population, time zone, latitude/longitude). So, as a result we can obtain a pretty good results using the Lithium NLP system. There is no direct comparison in the article with existing NLP systems, but on some aspects the Lithium NLP system is outperforms other systems. So, getting to conclusion about the second paper, we can say that it has potential in real world applications, which is applied on Lithium products. The best part of the paper is that it can conclude information on noisy inputs plus in multi-language inputs, which makes the task even harder, but it is where the world is shifting, since there is no common principle to follow. And even artificial languages are produced by human beings like Esperanto, but what waits us in the future is a mystery. The authors promise to work further, provide comparison with existing systems and work with academic and commercial systems.

The third article is about large scale relation extraction with reinforcement learning, which based on current solutions but slightly improves the results. Let’s check it out as well and see what the main difference versus current solutions is. The article states that previous solutions relied on manually labelled supervised data that is costly and limits to the number of relations and data size. So, Chinese researchers are trying to find a better solution to be able to apply for large scale data. As a possible solution they propose a novel model with reinforcement learning. The relation of the entities is used as distant supervision and guide the training relation extractor with the help of reinforcement learning methods. Also, in the paper the researchers are using two type of experiment – on small data and large one.

In the method, they introduce entity pair notion to identify how close two entities are. For instance, Steve Jobs and Apple. So, entity pairs are aligned into relations and the set of those relations form so called Bags. The relation extractor reads the Bags and outputs their extracted relations one by one. The researchers propose a novel method with reinforcement learning to learn relation extractor using large scaled distant supervised data. The bag relation is used as a distant supervision which guides the training of relation extractor. In general, the works mimics the previously implemented model called Distant Supervised Relation Extraction (DSRE), but slightly improves the result for every single sentence in the input dataset meaning that it has a potential and needs to be explored further more. Also, integrations with other neural network structures is possible.

In summing up all the three papers, we can say that the issue of Information Extraction is quite challenging and needs a lot of effort to be applied in order to find the most optimal solution. Also, we have learned that each model that is implemented and will be implemented should be aiming its target audience, since each field is complex enough and there almost no possibility to find a generic solution to all dataset in the world. We just covered medical and social media aspects, but there are lots of obstacles in each field, science, language and so on.

Machine Learning, Deep Learning and Artificial Intelligence is evolving and we will see where it takes us.